Sprachsynthese mit Mary

Simon Schmiedel

Proseminar „Tools für Computerlinguisten“

Was ist Sprachsynthese?

künstliche Erzeugung der menschlichen Sprache

Sprachsynthese

Sprachsynthese

Symbolische Repräsentation

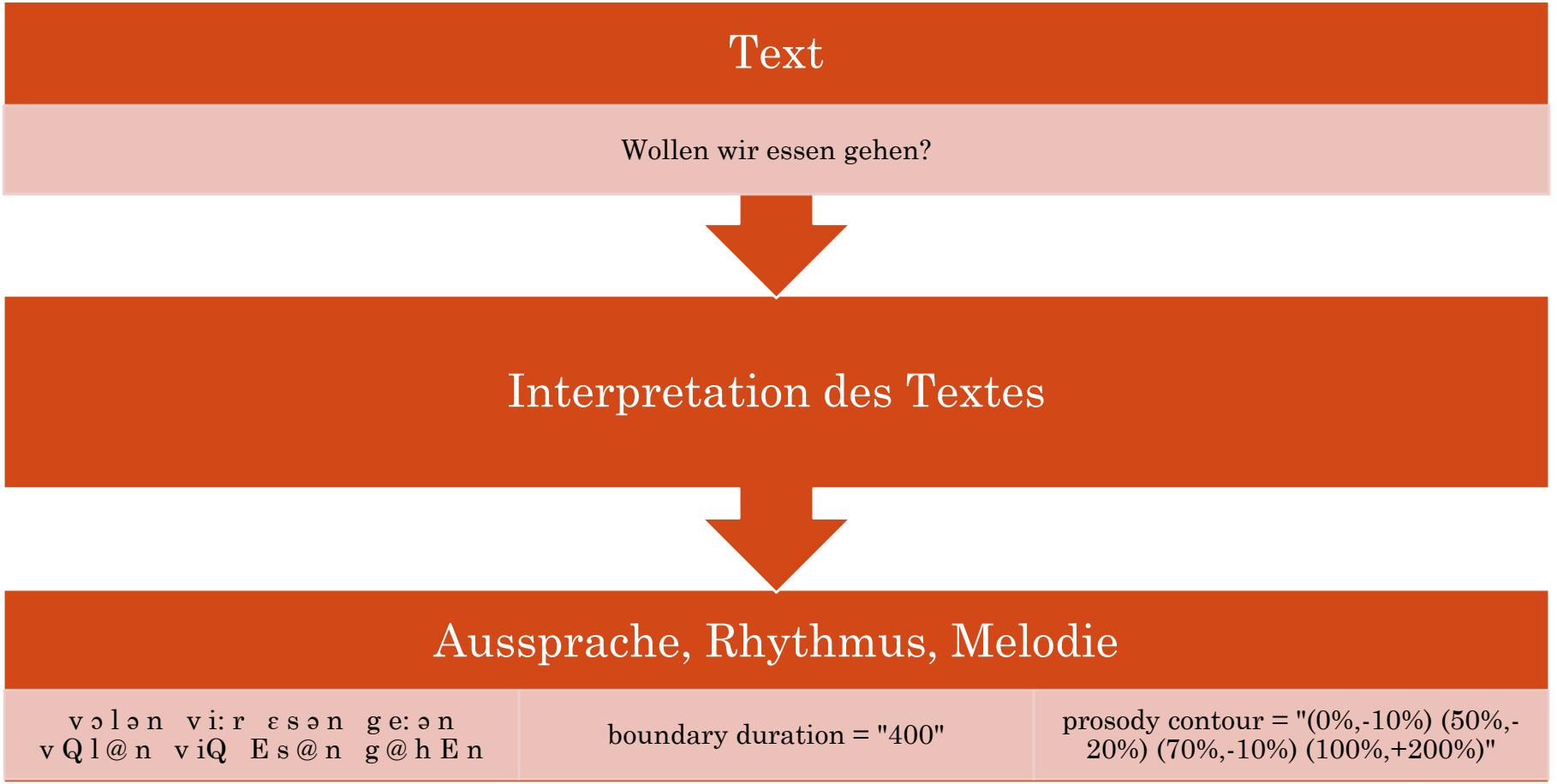
- Text-to-Speech:
  - Wollen wir zusammen essen gehen?
- Concept-to-Speech
  - Vorschlag(Sprecher, Hörer, „essen gehen“)

Akustische Repräsentation

Einsatzgebiete

- Mensch-Maschine-Kommunikation
  - Telekommunikationsdienste
- Hilfe für Behinderte
  - Screen Reader
- Fremdsprachenerwerb
  - Vokabeln vorsprechen lassen

Aufbau


Text
Wollen wir essen gehen?
Interpretation des Textes
Aussprache, Rhythmus, Melodie
v ɔ l ə n v i: r ɛ s ə n g e: ə n
v Q l @ n v iQ E s @ n g @ h E n
boundary duration = "400"
prosody contour = "(0%,-10%) (50%,-20%) (70%,-10%) (100%,+200%)"

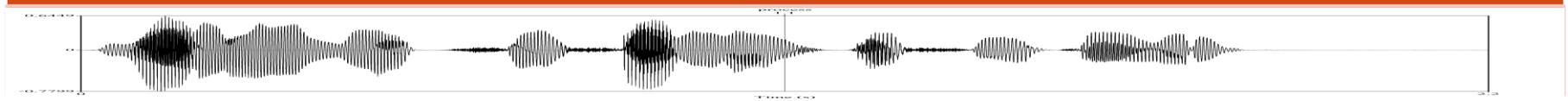
Aufbau

Aussprache, Rhythmus, Melodie

v ɔ l ə n v i: r ɛ s ə n g e: ə n v Q l @ n v iQ E s @ n g @ h E n	boundary duration = "400"	prosody contour = "(0%,-10%) (50%,-20%) (70%,-10%) (100%,+200%)"

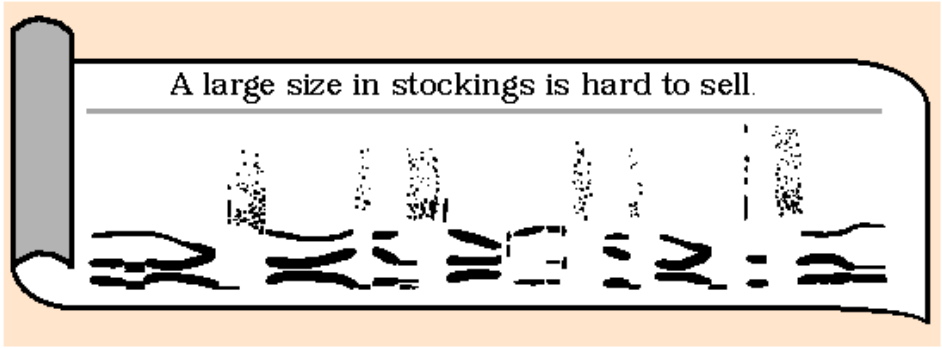
Sprachsignalgenerierung

Akustisches Signal



Formantsynthese

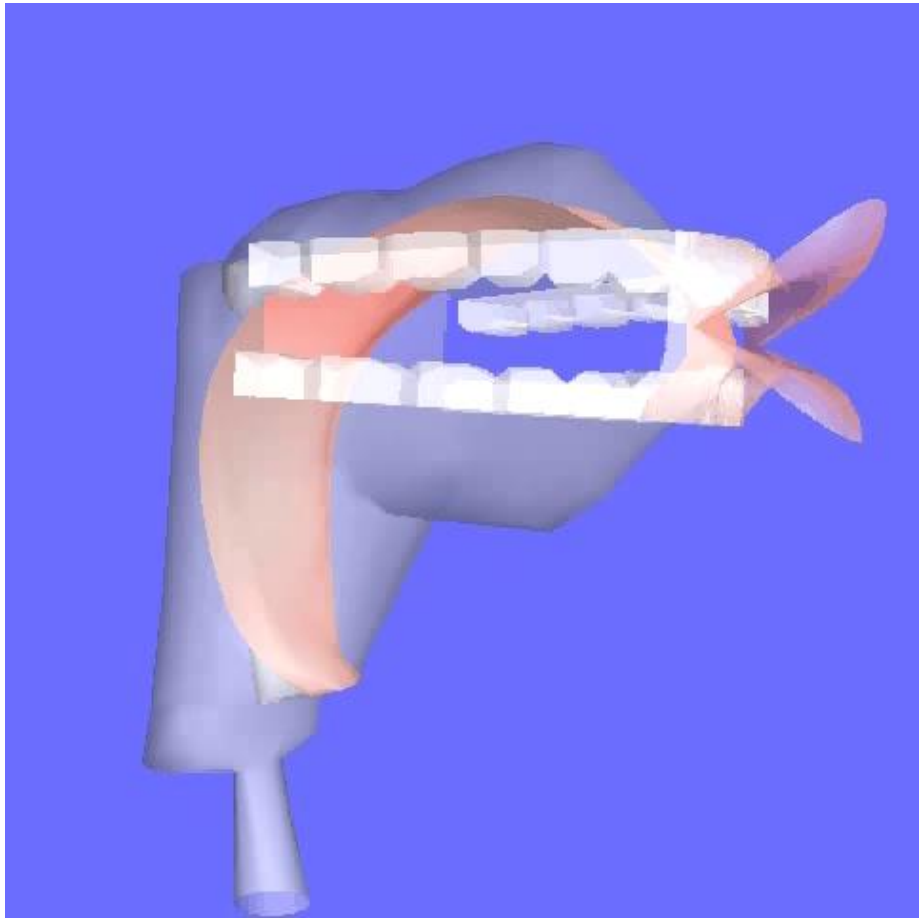
- **Wie hört sich menschliche Sprache an?**
- Additive Synthese
  - Klang durch Zusammenstellen der gewünschten harmonischen Teiltöne erstellt
- Parameter:
  - Grundfrequenz, Lautstärke, Phonation



Audio/Bild:
https://www.phonetik.uni-muenchen.de/studium/skripten/Haskins/Haskins/MISC/PP/SENTENCES/stockings.html

artikulatorische Synthese

- **<u>Wie produziert der Mensch Sprache?</u>**
- auf menschlichem Vokaltrakt und menschliche Artikulationsprozesse basierend



Video:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3628899/bin/pone.0060603.s008.avi

konkatenative Synthese

- Konkatenation natürlicher Sprachdaten
- Großes natürlich-sprachliches Korpus
  - Techniken:
    - Unit Selection
      - Segmente mit Akustische, phonetischer Eigenschaften
        - Grundfrequenzverlauf, Dauer oder Nachbarn.
      - Je größer der Korpus und vorhandene Segmente um so natürlicher die Synthese
        - Weniger Schnittstellen
    - Diphonsynthese
      - Datenbank von Lautübergängen
        - Lautmitte bis Lautmitte
      - Umfangreichere Datenbanken enthalten weitere Informationen zur Koartikulation

MaryTTS

Mary

- **M**odular **A**rchitecture for **R**esearch in **S**ynthesis.
- Beste OpenSource Text-to-Speech-Software
- Entwickelt am DFKI zusammen mit der Phonetik der UdS
- Verwendet die konkatenative Synthese

Marys Funktionsweise

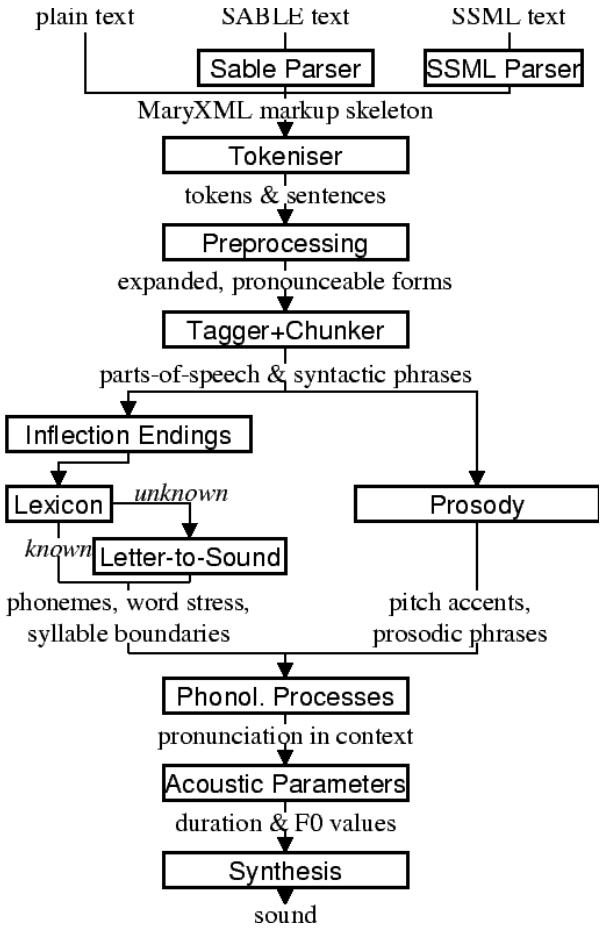


Bild: http://mary.dfki.de/documentation/module-architecture.html

- Jeder Schritt kann ausgegeben werden
- Mittels MaryXML kann in jeden Schritt eingegriffen werden

Marys XML

- Prinzipiell automatisch generiert
- Bietet dem User direkte Eingriffsmöglichkeiten
- Basiert lose auf SSML („Speech Synthesis Markup Language Version 1.0“, 2004)
  - Amazons Alexa verwendet SSML

Marys XML - wichtigste Tags

- <prosody>
  - Zur Modulierung von Grundfrequenz, Geschwindigkeit, Lautstärke
  - <prosody rate="-50%">
- <voice>
  - Ändern der Stimme
  - <voice name="bits3-hsmm">
- Besondere Stimme: Pavoque Styles
  - Unterstützt 4 Emotionen rudimentär (Sad, Happy, Angry, Poker)
  - <prosody style="angry">

Marys XML - Zwei Beispiele

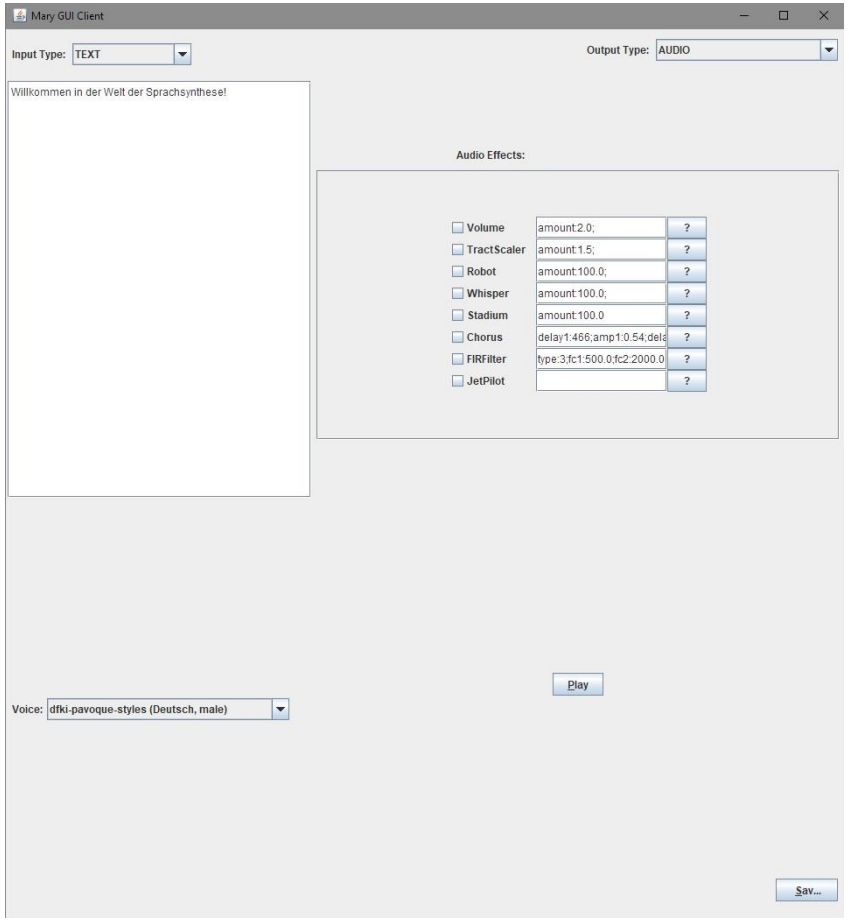
Prosody

```
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="en-US">

<prosody rate="-50%" pitch="+20%" range="-10%" volume="loud">
This is something you have to see!
</prosody>

</maryxml>
```

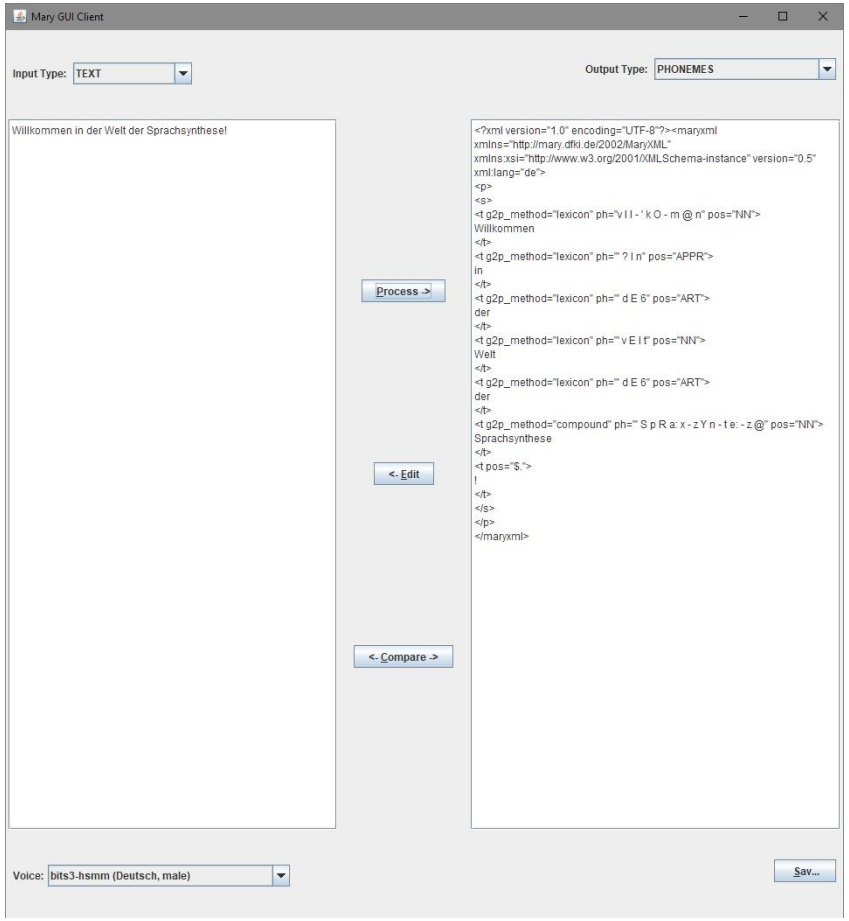
Letter-to-Sound

```
<?xml version="1.0" encoding="UTF-8" ?>
<maryxml version="0.4"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML"
xml:lang="de">

Klicken Sie den
<t ph="'ba-t@n">Button</t>,
um die
<t ph="'bi:-t@ls">Beatles</t>
zu spielen.

</maryxml>
```

Marys Client



Marys Client



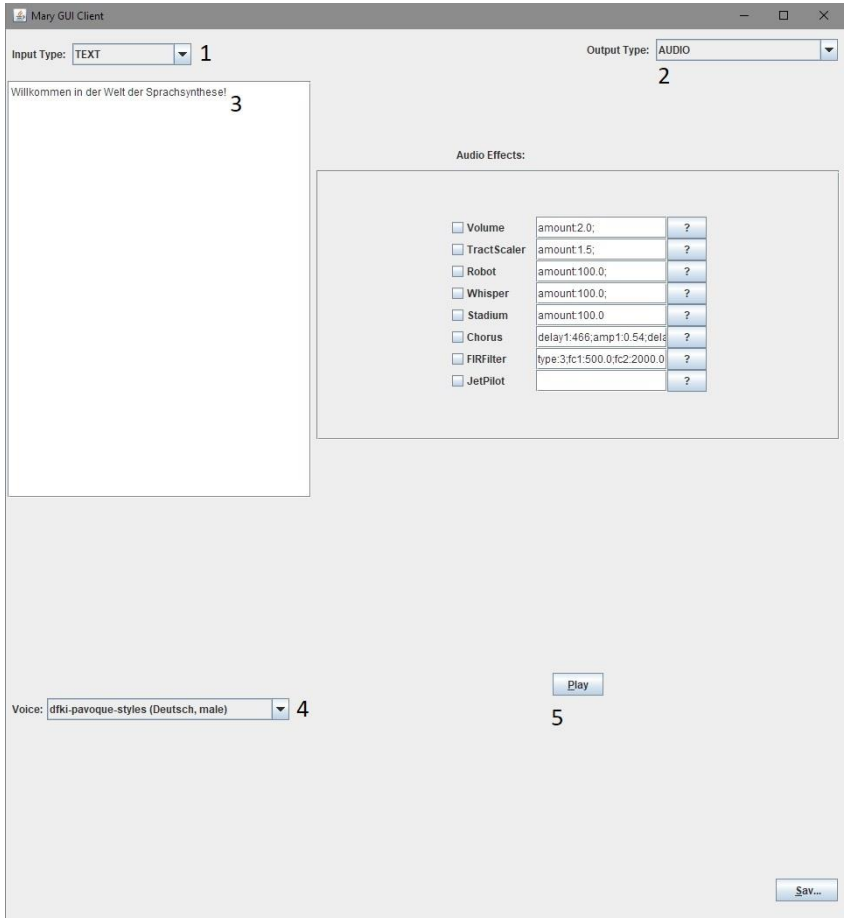
Mary ohne Client

- HTTP-Request an den Server
- Informationsabfrage
  - Version: *Server:Port*/version
  - Stimmen: *Server:Port*/voices
  - Sprachen: *Server:Port*/locales
- Synthese-Request
  - *Server:Port*/process?INPUT_TEXT=**Hello+world**&INPUT_TYPE=**TEXT**&OUTPUT_TYPE=**AUDIO**&AUDIO=**WAVE_FILE**&LOCALE=**en_US**
  - Weitere Parameter optional:
    - Audioeffekte
    - Stimme
    - Style
    - etc

Übungen

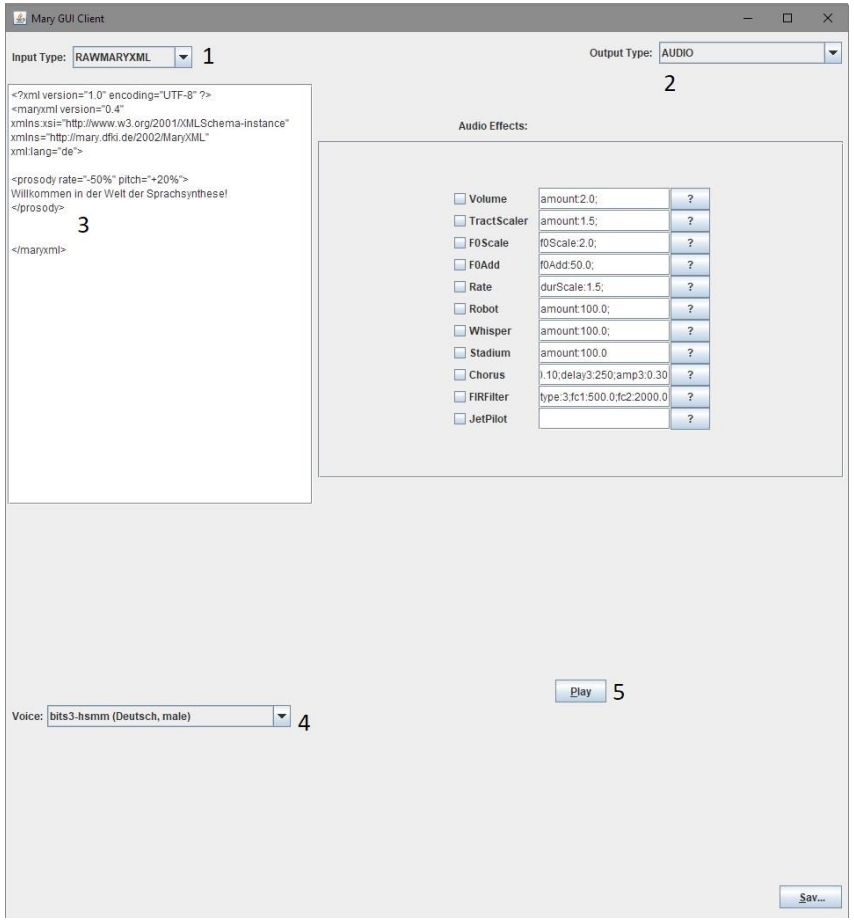
1. Aufgabe: Kennenlernen

1. Auf „Text“ als Eingabe stellen
2. Auf „Audio“ als Ausgabe stellen
3. Text eingeben
4. Stimme wählen
5. Play drücken



2. Aufgabe: MaryXML

1. Auf „RawMaryXML“ als Eingabe stellen
2. Auf „Audio“ als Ausgabe stellen
3. Text eingeben
4. Stimme wählen
5. Play drücken



2. Aufgabe: MaryXML -Tags

- Voice:
  - name: <voice name=„bits3-hsmm“>
- Prosody:
  - pitch: (Grundfrequenz)
    - <prosody pitch=„xHz“>
    - <prosody pitch=„+x%“>
    - <prosody pitch=„x%“>
  - rate: (Geschwindigkeit)
    - <prosody pitch=„+x%“>
    - <prosody pitch=„x%“>
  - range: (Variabilität des Pitches)
    - <prosody range =„xHz“>

- X = “x", “x.", ".x" und “x.x“ mit “x" als eine Folge von Ziffern
- Style:
  - Nur bei der Stimme „dfki-pavoque-styles“
  - <prosody style=„x“>
    - X= angry, happy, neutral, poker, sad
- Jedes geöffntete Element muss geschlossen werden:
  - <voice> → </voice>
  - <prosody> → </prosody>

3. Aufgabe: Eingriff in die Synthese

Text:

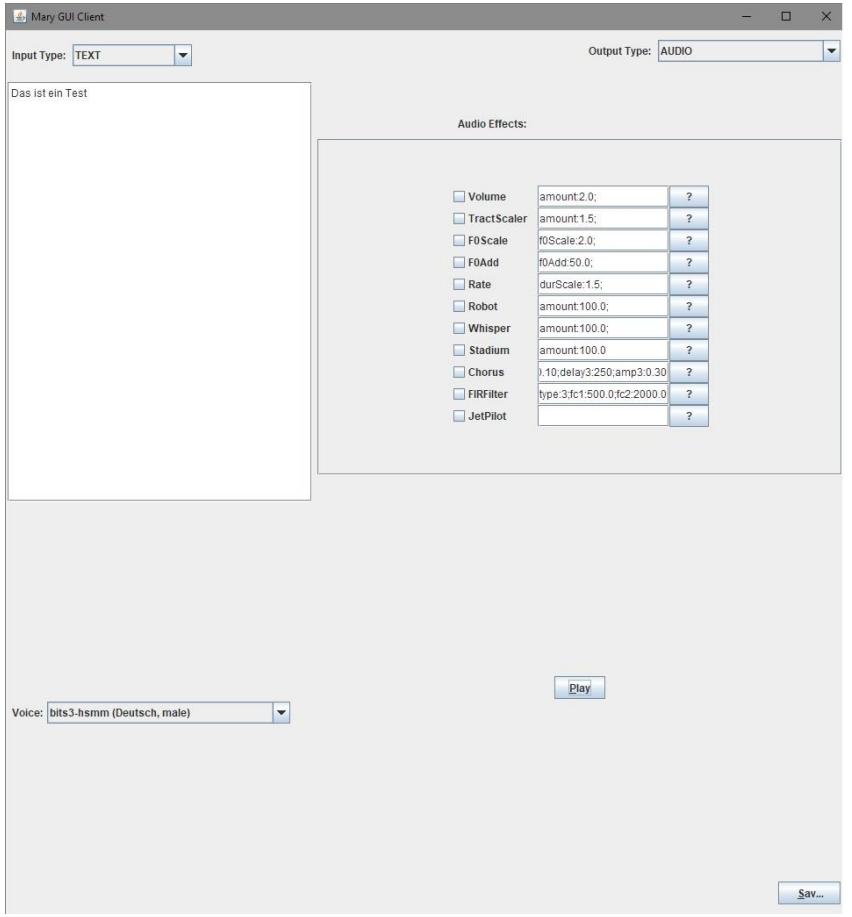
Klicke auf diesen Button

Stimme:

bits3-hsmm

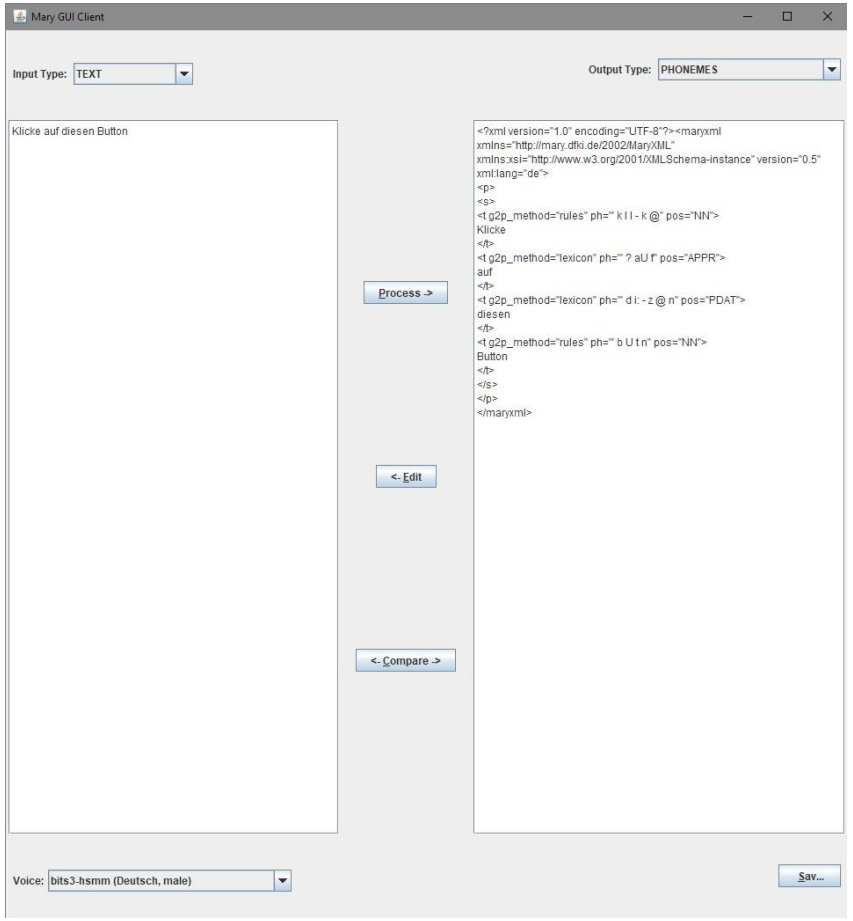
Abspielen

Was fällt auf?



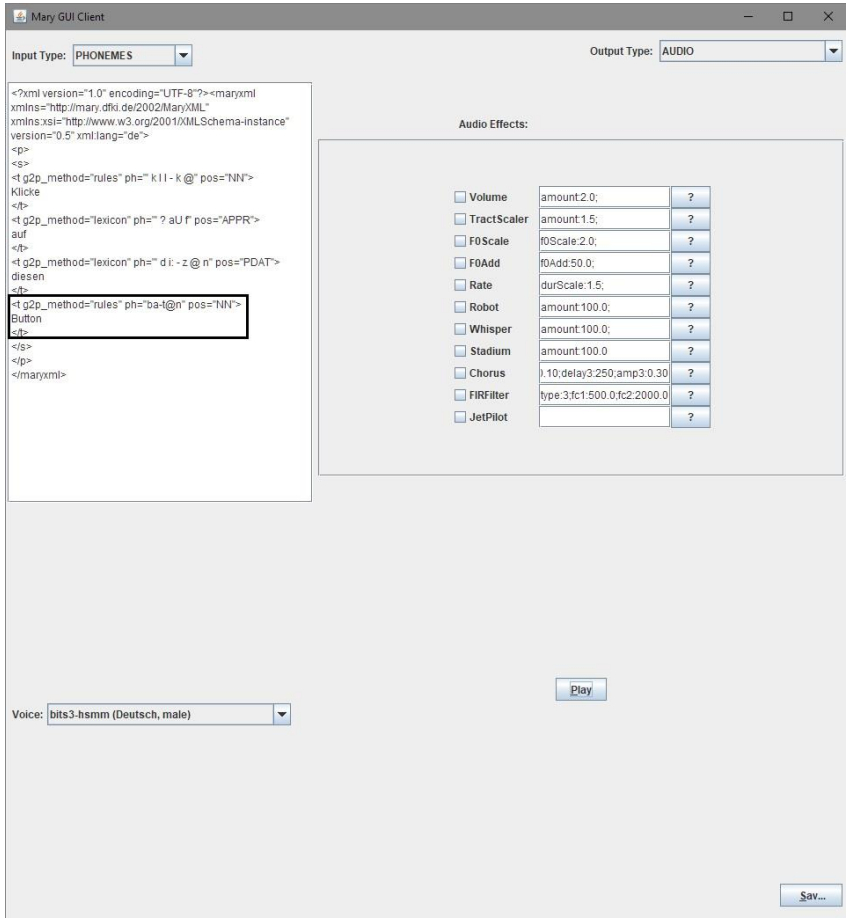
3. Aufgabe: Eingriff in die Synthese

1. Auf „Text“ als Eingabe stellen
2. Auf „Phonems“ als Ausgabe stellen
3. Text eingeben
4. Stimme wählen
5. Process drücken



3. Aufgabe: Eingriff in die Synthese

6. Auf „Edit“ klicken
7. „Button“ suchen
8. ph="b u t n" durch " ph="b a t @ n" ersetzen
   - Aussprache in SAMPA*
9. Auf „Audio“ als Ausgabe stellen
10. Play drücken



*http://www.phon.ucl.ac.uk/home/sampa/german.htm

Auf diese Weise kann prinzipiell in jeden Zwischenschritt eingegriffen werden

Zusammenfassung

Zusammenfassung

- Aufbau von Sprachsynthese
- Techniken Sprachsynthese
- Tool: MaryTTS

Danke für die Aufmerksamkeit